Predmet: **Mašinsko učenje 1**

Šema plana izrade projekta

Tema: **Procena vrednosti fudbalera sa Transfermarkt-a primenom regresionog modela**

Autori

Miloš Sirar IN 3/2020

Vasilije Zeković IN 4/2020

Datum: 30.01.2024.

Sadržaj

1 Uvod

U ovom delu projekta biće predstavljeni osnovni podaci o projektu i kratak plan kako će projekat biti razvijan. U dve glavne celine će biti objašnjeno kako smo došli do podataka, kako smo ih prilagodili za modele i koji regresioni modeli će biti upotrebljeni da bi se procenila vrednost fudbalera sa Transfermarkt-a.

Napomena: Uz ovaj plan je predat prethodno i fajl Odabir i analiza baze podataka, koji je dosta bitan da bi se razumeo ceo koncept projekta u nastavku. Informacije koje se tamo nalaze, neće se ovde ponavljati.

2 Istraživačka analiza podataka (EDA)

2.1 Postupak analize do konačne baze podataka

Nakon dodatnih analiza i razmišljanja, zaključili smo da baza podataka iz KT1 nije dovoljno dobra, odnosno da je moguće unaprediti je.

Na osnovu tih analiza, odlučili smo da neka od obeležja nisu značajna i da mogu biti izbačena iz skupa obeležja. Takođe, neka obeležja smo modifikovali u oblik koji je jednostavniji i smisleniji za tumačenje. Sada ćemo ukratko objasniti za svako od obeležja zašto je izbačeno, odnosno modifikovano.

1. date_of_birth
   - značajno obeležje
   - format ipak ne odgovara
   - generisano novo obežje iz njega -> old
   - ovo obeležje je obrisano iz skupa obeležja
2. citizienship
   - značajno obeležje

- format ipak ne odgovara
- generisano novo obežje iz njega -> citizienship_tier
- nacionalnost fudbalera predstavljena pomoću rangova na osnovu FIFA rang liste
- podeljeno u 5 kategorija (rangova), gde je 5 najbolja, a 1 najgora kategorija
- zašto? – vrednost fudbalera će biti veća za istu statistiku ukoliko je fudbaler po nacionalnosti u većim kategorijama
- ovo obeležje je obrisano iz skupa obeležja

Kategorija	Kategorija 5	Kategorija 4	Kategorija 3	Kategorija 2	Kategorija 1
Broj reprezentacija	10	15	15	30	34

3. club_domestic_competition_on_valuation_date
   - značajno obeležje
   - format ipak ne odgovara
   - generisano novo obeležje iz njega -> club_domestic_competition_on_valuation_date
   - podeljeno u 3 kategorije (ranga), gde je 3 najbolja, a 1 najgora kategorija
   - zašto? – vrednost fudbalera će biti veća za istu statistiku ukoliko je fudbaler po ligi u većoj kategoriji
   - ovo obeležje je obrisano

Kategorija	Kategorija 3	Kategorija 2	Kategorija 1
Domaće takmičenje	GB1, IT1, ES1, FR1, L1	PO1, RU1, TR1, NL1, BE1	URK1, SC1, GR1, DK1

4. valuation_date
   - kategoričko obeležje
   - značajno obeležje
   - sačuvano je jer je dosta bitno prilikom procene vrednosti fudbalera
   - dosta se razlikuje vrednost fudbalera za istu statistiku 2012. i 2023. godine (zbog fudbalskog tržišta i zbog inflacije)
   - svaka godina je podeljena na 3 dela (tertila)
   - ukupno imamo 35 tertila
   - svaka vrednost je predstavljena od 0 do 34, sa tim da je sa 0 predstavljen 2012-T2, a sa 34 2023-T3
   - ovo obeležje nije obrisano, samo se menja interpretacija podataka
5. player_id
   - nije relevantno za model
   - ovo obeležje je obrisano
6. first_name
   - nije relevantno za model
   - ovo obeležje je obrisano

7. last_name
   - nije relevantno za model
   - ovo obeležje je obrisano
8. agent_name
   - nije relevantno za model
   - pojavljuju se samo nazivi poslednjeg agenta fudbalera
   - to ne mora da znači da je fudber za sve ranije valuacije imao istog agenta
   - ovo obeležje je obrisano
9. contract_expiration_date
   - nije relevantno za model
   - pojavljuju se samo datumi isteka poslednjeg ugovora
   - to ne mora da znači da je tokom neke ranije valuacije važio taj ugovor
   - ovo obeležje je obrisano
10. current_club_name
    - nije relevantno za model
    - pojaviće se u jednoj od valuacija svakako kao ime kluba u trenutku valuacije
    - ovo obeležje je obrisano
11. current_club_domestic_competition
    - nije relevantno za model
    - isto kao i za current_club_name
    - ovo obeležje je obrisano
12. club_name_on_valuation_date
    - jeste bitan podatak, ali već smo napravili obeležje club_value_on_valuation_year
    - visoko je korelisano sa vrednošću kluba (nose istu informaciju)
    - po tržišnoj vrednosti možemo da procenimo koliko je dobar klub
    - nije nam potrebno ime kluba
    - ovo obeležje je obrisano

2.2 Konačna baza podataka

Nakon svih analiza i promena, došli smo do konačne baze podataka, koja ima naziv all_data_new_2.csv. U tabeli su prikazani nazivi obeležja iz baze, njihovi nazivi na srpskom i kratak opis svakog obeležja.

Attribute name	**Naziv obeležja**	**Opis obeležja**
valuation_date	Vrednost tertila datuma valuacije	*Vrednost tertila datuma valuacije na datum procene vrednosti fudbalera*

old	Godine fudbalera	*Broj godina fudbalera*
height_in_cm	Visina u cm	*Visina fudbalera izražena u cm*
citizienship_tier	Kategorija nacionalnosti	*Vrednost kategorije nacionalnosti fudbalera*
position	Pozicija	*Deo terena na kojem fudbaler igra (golman, odbrana, vezni red, napad)*
sub_position	Konkretna pozicija	*Tačna pozicija na kojoj fudbaler igra (primer: centralni vezni)*
foot	Jača noga	*Koja noga fudbalera je jača (primarna)*
market_value_on_valuation_date	Tržišna vrednost na datum valuacije	*Tržišna vrednost fudbalera u trenutku valuacije*
games_played_in_domestic_competition_on_valuation_date	Broj utakmica odigranih u domaćem takmičenju na datum valuacije	*Ukupan broj utakmica koje je fudbaler odigrao u domaćem takmičenju od trenutka prethodne valuacije do trenutka trenutne valuacije*
average_minutes_in_domestic_competition_on_valuation_date	Prosečan broj minuta u domaćem takmičenju na datum valuacije	*Prosečan broj minuta koje je fudbaler odigrao u domaćem takmičenju od trenutka prethodne valuacije do trenutka trenutne valuacije*
goals_in_domestic_competition_on_valuation_date	Broj golova u domaćem takmičenju na datum valuacije	*Ukupan broj golova koje je fudbaler postigao u domaćem takmičenju od trenutka prethodne valuacije do trenutka trenutne valuacije*
assists_in_domestic_competition_on_valuation_date	Broj asistencija u domaćem takmičenju na datum valuacije	*Ukupan broj asistencija koje je fudbaler ostvario u domaćem takmičenju od trenutka prethodne valuacije do trenutka trenutne valuacije*
yellow_cards_in_domestic_competition_on_valuation_date	Broj žutih kartona u domaćem takmičenju na datum valuacije	*Broj žutih kartona koje je fudbaler prikupio u domaćem takmičenju od trenutka prethodne valuacije do trenutka trenutne valuacije*
red_cards_in_domestic_competition_on_valuation_date	Broj crvenih kartona u domaćem takmičenju na datum valuacije	*Broj crvenih kartona koje je fudbaler prikupio u domaćem takmičenju od trenutka prethodne valuacije do trenutka trenutne valuacije*
games_played_in_european_competition_on_valuation_date	Broj utakmica odigranih u evropskim takmičenjima na datum valuacije	*Ukupan broj utakmica koje je fudbaler odigrao u evropskim takmičenjima od trenutka prethodne valuacije do trenutka trenutne valuacije*
average_minutes_in_european_competition_on_valuation_date	Prosečan broj minuta u evropskim takmičenjima na datum valuacije	*Prosečan broj minuta koje je fudbaler odigrao u evropskim takmičenjima od trenutka prethodne valuacije do trenutka trenutne valuacije*
goals_in_european_competition_on_valuation_date	Broj golova u evropskim takmičenjima na datum valuacije	*Ukupan broj golova koje je fudbaler postigao u evropskim takmičenjima od trenutka prethodne valuacije do trenutka trenutne valuacije*
assists_in_european_competition_on_valuation_date	Broj asistencija u evropskim takmičenjima na datum valuacije	*Ukupan broj asistencija koje je fudbaler ostvario u evropskim takmičenjima od trenutka prethodne valuacije do trenutka trenutne valuacije*
yellow_cards_in_european_competition_on_valuation_date	Broj žutih kartona u evropskim takmičenjima na datum valuacije	*Broj žutih kartona koje je fudbaler prikupio u evropskim takmičenjima od trenutka prethodne valuacije do trenutka trenutne valuacije*
red_cards_in_european_competition_on_valuation_date	Broj crvenih kartona u evropskim takmičenjima na datum valuacije	*Broj crvenih kartona koje je fudbaler prikupio u evropskim takmičenjima od trenutka prethodne valuacije do trenutka trenutne valuacije*
club_domestic_competition_on_valuation_date_tier	Kategorija domaćeg takmičenje kluba u trenutku valuacije	*Vrednost kategorije domaćeg takmičenja kluba u trenutku valuacije*
club_value_on_valuation_year	Vrednost kluba u godini valuacije	*Vrednost kluba u godini za klub za koji je fudbaler igrao u trenutku valuacije*

2.3 Modifikovani odgovori sa KT1

Kako smo ponovo analizirali i menjali bazu podataka, odgovori na neka pitanja sa KT1 su se promenila, pa ćemo ukratko da ažuriramo odgovore na njih.

6. Numeričkih obeležja ima 16.

7. Kategoričkih obeležja ima 6.

Kategoričko obeležje koje ima najmanje različitih kategorija je foot sa tri kategorije: right, left, both.

Kategoričko obeležje koje ima najviše različitih kategorija je sub_position, sa 13 različitih kategorija.

8. Obeležje koje se predviđa u našem modelu je: market_value_on_valuation_date. Opseg vrednosti za to obeležje je od 10 000 do 200 000 000. Prosek vrednosti je 9 646 029, dok je medijana 4 000 000.

9. Obeležja koja smo izbacili u odnosu na bazu podataka iz KT1 su detaljno objašnjena u tački 2.1.

11. Pojedini uzorci su izbačeni zbog političkih razloga.

13. Nakon svih sređivanja baze podataka, ostalo je 41 873 uzorka i 22 obeležja.

14. Većina numeričkih obeležja poseduje autlajere. Autlajeri svih obeležja predstavljaju ekstremne vrednosti, prirodne su i nisu posledica grešaka u merenju stoga ćemo ih zadržati.

15. Koeficijent asimetrije je 3.43, a koeficijent spljoštenosti je 18.

3 Obuka modela

3.1 Dummy varijable

Potrebno je transformisati preostala 3 kategorička obeležja u neki vid numeričkog obeležja. Za to smo koristili kreiranje dummy varijabli pomoću pd.get_dummies funkcije.

Svuda je korišten drop_first = True, pa je zato kreirano jedno manje obeležje nego što ima različitih vrednosti obeležja.

1. position – 4 pozicije: position_Defender, position_Goalkeeper, position_Midfield (position_Attack se ne kreira eksplicitno)
2. sub_position – 13 konkretnih pozicija: sub_position_Central Midfield, sub_position_Centre-Back, sub_position_Centre-Forward, sub_position_Defensive Midfield, sub_position_Goalkeeper, sub_position_Left Midfield, sub_position_Left Winger, sub_position_Left-Back, sub_position_Right Midfield, sub_position_Right Winger, sub_position_Right-Back, sub_position_Second Striker (sub_position_nan se ne kreira eksplicitno)
3. foot - Obeležje foot ima 3 moguće opcije: foot_left, foot_right (foot_both se ne kreira eksplicitno)

3.2 Podela skupa podataka

Obavezno ćemo uraditi podelu skupa na trening i test. Test će iznositi 10% skupa podataka. Zatim ćemo svaki model obučiti i oceniti koristeći unakrsnu validaciju trening skupa. Koristićemo mean absolute error pri oceni kvaliteta hiperparametara.

3.3 Algoritmi

Nakon što smo ovo odradili imamo sve neophodno da koristimo neki model. Ali pre toga moramo da se odlučimo koje algoritme ćemo da koristimo.

Algoritmi za koje smo se opredelili su:

1) Linearna regresija
2) KNN regresor
3) SVR – support vector regression

*Pored navedenih algoritama, probali smo još i RandomForestRegressor, XGBRegressor i neuronsku mrežu. Ali se nismo odlučili da li ćemo ih detaljnije ispitati i primeniti u našem projektu, ostaje da se vidi.

3.3.1 Linearna regresija

Prvo ćemo pričati o linearnoj regresiji. Odvojićemo vrednosti obeležja koje se predviđa od obeležja koja će služiti da se ta vrednost predvidi.

x – sadrži “market_value_on_valuation_date”, vrednost koju želimo da model predvidi

y – sadrži vrednosti svih ostalih obeležja, nakon transformacije sa dummy varijablama, ovde imamo 35 obeležja

Nakon što smo ovo uradili krećemo da ispitujemo različite modele linearne regresije da bismo videli kakve rezultate daju.

Probaćemo različite modele. Osnovnu hipotezu, zatim hipotezu sa i bez interakcija, kao i Ridge i Lasso regularizaciju. Koristićemo i standardizaciju. Takođe ćemo menjati stepen obeležja u osnovnoj hipotezi i ostale hiperparametre poput include_bias-a (uključivanje slobodnog člana), alpha (regularizacioni parametar) i fit_intercept-a.

3.3.2 KNN regresor

Sada ćemo pričati o KNN regresoru.

Kod ovog modela koristimo unakrsnu validaciju, kao i kod linearne regresije.

Probaćemo procenu hiperparametara modela u dva slučaja, sa I bez standardizcije.

Pomoću GridSearchCV ćemo pronaći koji su najbolji parametri.

Gledaćemo različite vrednosti za n_neighbors, a za metrike ćemo gledati: hamming, euclidean, manhattan, minkowski (za p različito od 1, 2 I beskonačno), chebyshev. Pretpostavljamo da će prvenstveno metrike poput euklidske i manhattan pokazati najbolje rezultate, neke metrike su stavljene čisto ekspirementalno.

Gledaćemo da se scoring radi na osnovu mean absolute error (kvalitet modela u pogledu objašnjene varijanse) uz adekvatan izbor parametara cv (koliko podskupova ima u unakrsnoj validaciji i koje je fiksno za sve algoritme).

Nakon uspešne selekcije najboljih hiperparametara (i u odnosu na PCA) primenićemo iste za obuku modela nad čitavim trening skupom, uz evaluaciju svih skorova dobijenih nad test skupom. Koristićemo KNeigborsRegressor da bismo napravili regresor. A od parametara ćemo inicijalizovati n_neighbors (broj suseda) i metric (metrika).

3.3.3 SVR – Support vector regression

Poslednja vrsta algoritama sa kojom ćemo raditi naše predviđanje jeste SVM (support vector machine) za rešavanje regresionih problema.

Kao I do sada, planiramo da primenimo unakrsnu validaciju nad trening skupom pri proceni hiperparametara koristeći GridSearchCV.

Obavezno ćemo primeniti i standardizaciju kako bismo ubrzali proces konvergencije, što ume biti značajno pri kombinovanju hiperparametara I evaluacije istih.

Fokusiraćemo se na sledeće hiperparametre koji su značajni kod SVR-a:

- kernel, probaćemo različite kernel funkcije, poput linear, poly, rbf, sigmoid i slično.
- C ćemo probati za različite vrednosti.
- degree ćemo menjati pri korišćenju polinomijalnog kernela,
- coef0, odnosno slobodan član za polinomijalni I sigmoidni kernel koji ćemo takođe proveriti kako se ponaša uz promenu vrednosti,
- gamma kao parametar rbf kernela ćemo takođe ispitati detaljnije

Nakon što smo utvrdili najbolje hiperparametre (i u odnosu na PCA), koristićemo ih za obuku modela nad čitavim trening skupom, i kasniju ocenu rezultata pomoću različitih skorova.

3.4 PCA – Principal component analysis

Nećemo koristiti linearnu diskriminantnu analizu (LDA), jer je u pitanju regresioni problem. Prilikom smanjivanja dimenzionalnosti koristićemo metodu razlaganja na glavne komponente (PCA).

Hiperparametar o kom ćemo voditi računa jeste n_components, probaćemo redukovati dimenzionalnost i sa odabirom tačno određenog broja obeležja (broj komponentni), ali i sa željenim udelom varijanse koja će biti očuvana i nakon redukcije. Cilj nam je da očuvamo varijansu iz podataka na što većem nivou, a da opet smanjimo dimenzionalnost. Pre primene PCA metode ćemo obavezno standardizovati obeležja.

*Obavezno ćemo ponoviti sve algoritme i modele sa skupom podataka čije su dimenzije redukovane nakon primene PCA metode. Test skup će ostati isti, iznosiće 10% skupa podataka. PCA će biti fit-ovana nad trening skupom, i kao takva korištena za transformaciju trening i test skupa. Nad trening skupom će za sve algoritme i modele biti primenjena unakrsna validacija, a nakon toga korištena. Takođe kao I do sada scoring će se raditi nad mean absolute error.

4 Prezentacija rezultata

Na kraju kada odradimo sve analize i prođemo kroz sve modele, slediće prezentacija dobijenih rezultata.

Sledi izbor po jednog najboljeg modela za svaki od tri algoritama i 6 mogućih slučajeva (tri kod PCA i tri kod neredukovanog skupa podataka). Zatim obavezna obuka nad čitavim trening skupom. Uz evaluaciju rezultata konačnog modela dobijenih korišćenjem test skupa. Za evaluaciju tri izabrana krajnja modela podrazumevano će se koristiti R kvadrat skor, srednja apsolutna greška, prilagođeni R kvadrat skor, srednja kvadratna greška i koren srednje kvadratne greške.